2024

“Was this answer helpful?” – A Taxonomy for Feedback Mechanisms in Customer Service Chatbots

Daniel Schloß
Karlsruhe Institute of Technology, Germany, daniel.schloss@kit.edu

Saskia Haug
Karlsruhe Institute of Technology, Germany, saskia.haug@kit.edu

Alexander Mädche
Karlsruhe Institute of Technology, Germany, alexander.maedche@kit.edu

“Was this answer helpful?” – A Taxonomy for Feedback Mechanisms in Customer Service Chatbots

Research Paper

Daniel Schloß[1], Saskia Haug[1], Alexander Mädche[1]

[1] Karlsruhe Institute of Technology, human-centered systems lab, Karlsruhe,
{daniel.schloss,saskia.haug,alexander.maedche}@kit.edu

Abstract. Chatbot technology has rapidly spread, especially in digital customer service. However, the automation potential of chatbots can only be realized if customers are satisfied with their service. Collecting explicit feedback is a promising technique for assessing customer satisfaction and identifying issues with the chatbot. It enables chatbot managers and developers to enhance performance and design of operational chatbots on an informed basis. The evident significance of explicit customer feedback comes with a multitude of design options available. However, there is a lack of research on chatbot feedback mechanisms and practical as well as theoretical clarity. In this paper, we adress this gap by introducing a chatbot feedback taxonomy derived from existing research and a sample of N = 72 real world customer service chatbots. Furthermore, based on a cluster analysis, we identify four archetypes of feedback mechanisms and provide strategic guidelines for the informed use of each of those feedback design variants.

Keywords: Customer Service, Chatbots, User Feedback, Customer Feedback

1 Introduction

Chatbots have become an integral part of the customer service repertoire of organizations and institutions (De Keyser et al., 2019), accelerated by the recent technological advancements and popularity of the GPT models. For practitioners, such as customer service managers, who strive for a satisfactory but efficient digital service offering, it is essential to continuously monitor chatbots in operation and identify opportunities for improvement (Beaver and Mueen, 2020; Lewandowski et al., 2022). Beyond log data like conversational transcripts, leveraging explicit customer feedback is of particular value (Akhtar et al., 2019; Van Oordt and Guzman, 2021). Requesting feedback allows customers to articulate perceptions and opinions that might have remained hidden in the basic log data. With explicit feedback, customers can provide ratings, highlight issues, and propose improvements. Consequently, it is advisable to solicit feedback explicitly from customers using feedback mechanisms within customer service chatbots (Lewandowski et al., 2022; Van Oordt and Guzman, 2021). However, the design of feedback mechanisms in chatbots has hardly been researched compared to the non-conversational web (e.g., Fricker and Schonlau, 2002). Furthermore, explicit feedback

19th International Conference on Wirtschaftsinformatik,
September 2024, Würzburg, Germany

mechanisms have particular characteristics, as the feedback can be embedded variously in the conversation. For example, the feedback wording could be more response-oriented ("Is this answer helpful?") or service-oriented ("How satisfied were you with my service?"). Other factors such as timing, initiation or survey design also vary. Yet there are always advantages and disadvantages associated with each feedback mechanism. One design may have a greater impact on the customer's user experience, while another may only lead to a smaller amount of feedback (Akhtar et al., 2019). For this reason, in addition to usability heuristics and findings from general feedback research, decision-makers benefit from an overview and conceptual clarity (Nielsen, 1994; Tizard et al., 2020). However, yet, there is no framework for chatbot feedback mechanisms though feedback can be valuable to understand and improve the customer experience and service quality (Akhtar et al., 2019; Pagano and Bruegge, 2013; Xiao et al., 2021). For this reason, in this paper, we aim to answer the following research questions (RQ):

RQ1: Which explicit feedback mechanisms exist for customer service chatbots?
RQ2: How and according to which criteria can they be classified in a taxonomy?

Answering the RQ, in this paper we provide a taxonomy on feedback mechanisms for the collection of explicit feedback in customer service that is not only theoretically derived but also practically validated. To do so, we began reviewing related research on chatbots and user feedback (chapter 2). We then outlined the methodology for our taxonomy development project (chapter 3) (Nickerson et al., 2013). Since chatbot feedback already is practically established but theoretically unaddressed, we chose an inductive approach for taxonomy development, collecting a sample of N = 72 customer service chatbots, which served as empirical data basis (chapter 4.1) (Nickerson et al., 2013). Then, we iteratively developed the taxonomy by gradually adding and coding samples (chapter 4.2). We applied a clustering method to our taxonomy to identify relevant clusters in our sample (chapter 4.3). Ultimately, we were able to identify 4 chatbot feedback archetypes which we named and discussed based on existing research with regard to their characteristics and suitable applications (chapter 4.4). This paper concludes this study with a discussion, limitations and and a research outlook (chapter 5).

2 Related Work

2.1 Customer Service Chatbots

Digital customer service is increasingly supported by automation through web-based chatbots, which have a task-oriented design (De Keyser et al., 2019; Schuetzler et al., 2021). They can be found at the websites of service providers as in telecommunication, energy, finance, banking, insurance and mobility industry (Adamopoulou and Moussiades, 2020). When users seek assistance from those, they expect precise and reliable information or transactions and ultimately task fullfillment (Brandtzaeg and Følstad, 2017; Grudin and Jacques, 2019; Kvale et al., 2021). For this purpose, an established approach is user intent classification (of the user request) via Natural Language Understanding (NLU) followed by an answer which is either retrieved from a database or

generated via Natural Language Generation (NLG). Experts predict that hybrid models consisting of controllable NLU and NLG elements will become the standard in highly specialised contexts like customer service applications (Greyling, 2023; Von Straußenburg and Wolters, 2023). In addition to the natural language processing (NLP), customer experience is also shaped by dialog or frontend design or backend performance (Gao et al., 2021; Li et al., 2020; Zierau et al., 2020). If it is negative, it casts a negative light on chatbot technology as well as the associated company (Meyer-Waarden et al., 2020). Therefore, researchers have pointed out the importance of quality testing and criteria for chatbots early on (Maroengsit et al., 2019; Janssen et al., 2021). However, the majority of existing evaluation concepts, involving both experts and test users, focuses on the pre-deployment phase and controlled environments (Maroengsit et al., 2019). Due to their widespread adoption from 2016 onward, however, chatbots are required to demonstrate their effectiveness in daily operation (Dale, 2016; Følstad and Taylor, 2021). This can be supported by collecting explicit chatbot feedback in the field. On the one hand, it indicates performance and satisfaction, on the other hand, it is a pointer for weaknesses in the chatbot (Pagano and Bruegge, 2013; Akhtar et al., 2019; Kvale et al., 2021). Since continuous monitoring and updating of chatbots is a critical success factor, chatbot feedback becomes part of product development (Janssen et al., 2021; Lewandowski et al., 2022), either asynchronously or directly feedbacking AI training in the case of NLG chatbots (OpenAI, 2023; Ricciardelli and Biswas, 2019).

2.2 Customer Feedback for Customer Service Chatbots

In user feedback research, feedback is categorized according to explicitness, structure, initiation and valence. First, users or customers can provide *explicit* feedback by participating in surveys or reviews. *Implicitly*, they provide feedback through their actions, for example clicks or view times (Poblete and Baeza-Yates, 2008; Ordenes et al., 2014). In the case of chatbots, implicit feedback is available in the form of conversation log data, while our study focuses on explicit feedback which is often *structured* with (numerical) scales, thumbs or stars combined with an option for *unstructured* text input. While unstructured feedback is more time-consuming to analyze, it can offer greater insights for companies (Witell et al., 2011; Ordenes et al., 2014). Regarding *initiation* chatbot feedback provision may be a *passive* option (like a menu item), or the user is *actively* asked for feedback ("Was this response helpful?") (Wirtz and Tomlin, 2000). Generally, active requests for feedback are often linked to a specific event, for example after a process has been completed ("How was your call quality?") and is known to have higher response rates than passive feedback, but interrupts customers in their task (Sampson, 1998; Tizard et al., 2020). Lastly, regarding valence, feedback can be negative, neutral or positive (Van Doorn et al., 2010). Unsatisfied customers are often more inclined to give feedback, which leads to a bias, yet negative feedback holds particular value (Nasr et al., 2014; Akhtar et al., 2019). In chatbot literature, customer feedback is mentioned in the context of chatbot evaluation, but dedicated research on chatbot feedback design is scarce. However, it is a well-known problem that feedback in chatbots is rarely provided; response rates of 5% (passive) to 18% (active) have been observed (Akhtar et al., 2019; Kvale et al., 2021). In the case of customer service chatbots,

the incentives to provide feedback are low, since there is no opportunity to help others and no long-term usage (Hennig-Thurau et al., 2004; Xiao et al., 2021). Nevertheless, customer service chatbots are well-suited for collecting customer feedback. On the one hand, they can leverage "in-app feedback mechanisms" and obtain situational feedback (Van Oordt and Guzman, 2021). On the other hand, studies with dedicated feedback or interview chatbots have shown that requesting feedback via conversational interfaces has an engaging effect on users (Te Pas et al., 2020; Han et al., 2021; Xiao et al., 2021).

3 Method

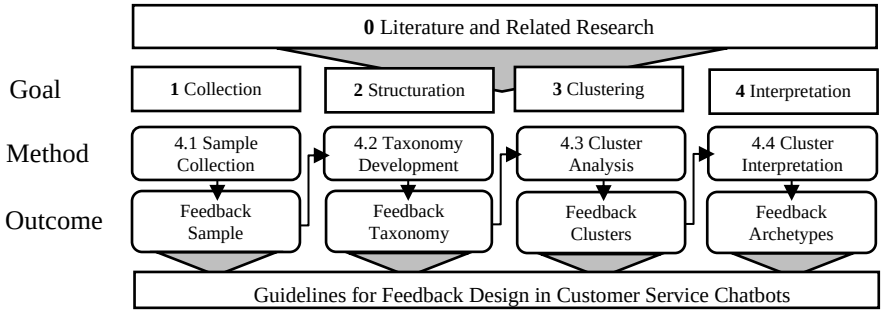

Figure 1. Research Method

The objective of this study was to develop a taxonomy for mechanisms to collect explicit feedback in customer service chatbots. Taxonomies can help to structure and organize the body of knowledge in a field, enabling researchers to study relationships and concepts (Glass and Vessey, 1995; Nickerson et al., 2013). As **Figure 1** shows, we first examined related research literature (0). We then collected an empirical market sample consisting of 72 customer service chatbots (with 84 feedback mechanismus) of seven german service-oriented industries (1). Following an established methodology for taxonomy building (Nickerson et al., 2013), we set meta-characteristics and end conditions and iteratively developed dimensions and characteristics for the feedback mechanisms by repeatedly drawing a subset of our sample (2) (Strauss and Corbin, 1998; Glaser and Strauss, 1999). After the taxonomy was finalized, we statistically identified 4 clusters (3). In the next step, we defined and named the 4 general chatbot feedback archetypes we found (4). Finally, we discussed these archetypes regarding their suitability and applicability, deriving guidelines for feedback collection in customer service chatbots.

4 A Taxonomy for Chatbot Feedback Mechanisms

4.1 Sample Collection

Having defined our research problem and goal, we collected chatbots with feedback mechanisms to form the sample of our targeted taxonomy. We focused on german service-oriented industries, namely energy and utilities, public administration, insurance,

mobility, finance and banking, and telecommunications where the use of chatbots is typically worthwhile (Fitzsimmons and Fitzsimmons, 2011; Kvale et al., 2021). We found a total of 79 chatbots across more than 200 website visits based on company registers. We excluded 7 chatbots which did not have any feedback mechanism. Of the remaining 72 chatbots, 34 were from energy and utilities, 11 from public administration, 11 from insurance companies, 7 from mobility and travel providers, 5 from financial institutions and 3 from the telecommunications sector. This distribution reflects the general presence of market participants, e.g. there are many more energy suppliers than telecommunications providers in germany and we do not expect any correlation between industry and feedback mechanisms. All companies are service providers that process typical customer service issues such as FAQ, bookings or cancellations (De Keyser et al., 2019). Technologically, all of the chatbots in our sample used some kind of NLU-based intent classification and response retrieval, which was the market standard at the time of the sample collection (2023). All of them were supplemented with buttons for navigation and a partion included pre-defined multi step dialogs (23) or click paths (12, navigation without free text input). When collecting the sample, we noticed two interesting aspects: When we looked for the chatbot vendors, directly or via the embed code, we noticed that some companies had the same chatbot vendor or platform. This entails the risk of an imbalance in the sample and taxonomy (Strauss and Corbin, 1998; Lin et al., 2017). However, we found that, there were no major clusters, as the largest group were custom chatbots built in-house, see **Figure 2**. Furthermore, we saw a broad spread of providers, a total of 24 plus 16 custom chatbots.

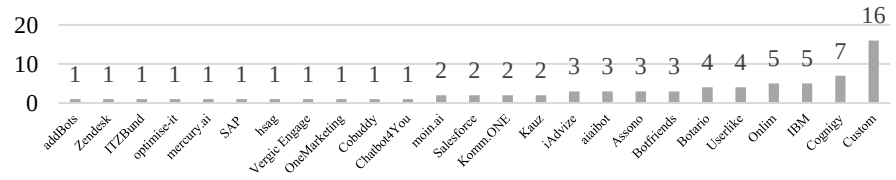

Figure 2. Chatbot Platforms and Vendors in the Chatbot Sample

We could also detect that some chatbot platforms such as Cognigy offer several feedback mechanisms via in-house development, for example regarding feedback dialogs (Strohmann et al., 2023). Additionally, some of the 72 chatbots had even more than one feedback mechanism within one interface. For example, one chatbot had a passive feedback item in the header of the chatbot interface but also actively asked customers whether their question had been answered. This led to a total of 84 samples to proceed.

4.2 Taxonomy Development

Building on Nickerson et al. (2013), the first step in our taxonomy building was determining a meta-characteristic for the taxonomy, which is the basis for choosing *dimensions* and *characteristics* to include in the taxonomy. Our meta-characteristic to observe were the "feedback mechanisms in customer service chatbots." Next, we defined the following ending conditions under which the taxonomy development terminates:

- all chatbots in the sample have been examined
- at least one chatbot is classified under each characteristic of each dimension
- no new dimensions or characteristics were added, merged or split in the last iteration
- every dimension is unique and not repeated
- every characteristic is unique and not repeated within its dimension

Having determined the ending conditions, there are two options to proceed. Researchers can decide to choose an empirical-to-conceptual, i.e., inductive, or an conceptual-to-empirical, i.e., deductive approach, depending on their data and domain knowledge (Nickerson et al., 2013). Since we intended to use empirical samples to work inductively, we chose the empirical-to-conceptual approach in all our iterations. In this approach, a subset of samples that should be classified is chosen. This subset can be of random, convenient or systematic size (Nickerson et al., 2013). We chose subsets of 2, 2, 1, 1, 1, 1, 11, 31 and 22 chatbots for the respective first to ninth iteration, which illustrates the gradual theoretical saturation (Glaser and Strauss, 1999). In each iteration, we draw a subset of the feedback mechanisms and analyzed those individually among the three authors. The subsets were labeled and grouped independently and dimensions and characteristics in each dimension were individually defined. The assignment was then discussed: If a feedback mechanism was labeled differently or there were differences in the dimensions or characteristics, these were discussed in order to decide on a categorization for each mechanism and a final taxonomy variant. In order to avoid ambiguities regarding the preliminary taxonomy in any iteration, a coding guide was created and updated, detailing dimensions, characteristics and their differences with screenshots and examples. At the end of each iteration, the end conditions were checked. If they were not met, another iteration began, resulting in a new version of the taxonomy, possibly with dimensions or characteristics updated. After nine iterations, no dimensions or characteristics were added or altered, at least one object was classified under each characteristic of each dimension, and each of those characteristics and dimensions were unique and balanced (Nickerson et al., 2013).

The final taxonomy (see **Table 1**, next page) consists of the dimensions *Feedback Initiation*, *Feedback Access*, *Feedback Presentation*, *Feedback Structure*, *Pre-Labeling*, *Feedback Specificity* and *Feedback Scope*. The characteristics of all dimensions are exclusive, i.e. each chatbot feedback mechanism can only have one characteristic. The only exception is the *Feedback Structure*, as characteristics can be combined, e.g. when a scale and a free text field are used. **Table 1** also shows at which iteration a dimension or characteristic was added (blue columns). Some characteristics were added while drawing further samples, others were formed on the basis of a differentiation of characteristics discussed by the authors. The first dimensions were created based on the user feedback literature (e.g. active vs. passive or feedback structure) and discussion of the authors guided by the overall sample data (e.g. Wirtz and Tomlin, 2000).

The first dimension, *Feedback Initiation*, captures if feedback is actively or passively elicited (Sampson, 1996). Active initiation implies that the chatbot prompts the customer to provide feedback. This can be implemented within the chat or when closing the chatbot interface, for example. With passive initiation, the customer initiates the feedback e.g. by asking for a feedback option or clicking on UI elements. *Feedback*

Access describes the situations in which a user can access the chatbot's feedback mechanism(s). The most common feedback access was the automatic display at the end of the conversation. This encompasses scenarios where the user him- or herself closes the chatbot as well as after predefined end events or timeouts. Another very common characteristic were feedback prompts after scripted multi-step dialogs. In this case, the feedback appears after completing a specific task, e.g. ordering a new credit card, allowing the user to provide feedback after ending the task dialog while the conversation can still continue (Følstad and Taylor, 2021). The characteristic Top-level Item refers to a UI element located prominently at the top of the chatbot user interface, such as in the header bar, where feedback is accessed by clicking on the item. Similarly, the characteristic Menu Item describes a passive UI element, which can be accessed through a dropdown menu or similar UI elements. In-chat buttons after response are characteristic buttons (usually thumbs up/thumbs down) that are visible in the chat itself and usually located close to the text bubbles. They are usually very close and related to responses of the chatbot. Lastly, the characteristic Intent refers to feedback access via NLU-classification of messages like "i have a complaint".

Table 1. Final Taxonomy

Dimension	Iteration	Characteristic	Iteration	N
Feedback Initiation	1	Active	1	44
		Passive	1	38
Feedback Access	1	End of conversation	3	23
		End of dialog	3	21
		Top-Level Item	8	15
		In-chat buttons after response	4	10
		Intent	2	10
		Menu Item	8	3
Feedback Presentation	2	Pop-Up Window	2	28
		In-chat buttons	6	26
		External Link	2	14
		Chat Widget	5	8
		Dialog	2	6
Feedback Structure	1	Scale	1	73
		Options	1	8
		Free Text	1	47
		Unstructured	2	8
Pre-Labeling	1	Pre-Labeling	1	61
		No Pre-Labeling	1	21
Feedback Specificity	4	Service Delivery	4	56
		Message Content	4	10
		Company	4	9
		Technology	8	1
		Unspecified	7	6
Feedback Scope	3	Conversation	3	44
		Dialog	3	20
		Single Response	4	10
		General	3	8

Feedback presentation characterizes the way the feedback items or surveys are presented to the customer. The characteristic Pop-Up Window refers to an overlay window within the chatbot covering the message interface. In-chat buttons are buttons directly below or on chat messages, while Chat widgets represent cards and forms located in between the chat messages. The characteristic Dialog implies the presentation of the feedback feature as a dialog, with the feedback survey having the same appearance as

the normal chat. Lastly, the characteristic External Link denotes that the feedback feature is opened in a new browser tab or website.While the initial three dimensions primarily address the invocation and usability of feedback mechanisms in chatbots, the subsequent dimensions focus on the information captured and its context:
The *Feedback Structure* dimension refers to the form in which user feedback is captured, encompassing characteristics such as Scale (rating statements on a scale with two to n values), Options (choosing from non-numeric alternatives), Free text (providing textual input based on survey questions), and Unstructured (feedback without without specific survey questions). The *Pre-Labeling* dimension is binary and indicates whether the feedback query uses a form of pre-classification, e.g. complaint/suggestion or "I was not understood correctly" and "The information was insufficient". *Feedback Specificity* describes if the subject of the feedback is specified by the service provider. Given the formulation of feedback questions (e.g. "Do you like our service?") or feedback access, the feedback can relate more strongly to a single Message, overall Service Delivery, the chatbot Technology or the service Company. In instances where the relationship between feedback and these categories was undefined, the characteristic Unspecified was assigned. Conversely, the *Feedback Scope* dimension addresses the parts of the chatbot interaction that users refer to with their feedback. They may refer to the entire Conversation, specific Dialogs, Single Responses, or they may provide General Feedback, extending beyond the conversation. This broader feedback may prompt users to share their opinions on their overall service interaction with the company.

4.3 Cluster Analysis

The dimensions and characteristics found in our taxonomy can be used to describe the feedback collection in web-based customer service chatbots. However, each feedback mechanism is only described by the sum of its characteristics. Moreover, certain dimensions, for example feedback access and feedback scope, relate to each other. Therefore, the goal of this step was to find groups within the feedback sample that describe common combinations of the characteristics. If these can be determined, it is possible to abstract archetypes that can be used to structure the design knowledge on chatbot feedback (Glass and Vessey, 1995). In order to find those, a data-driven cluster analysis was conducted to form clusters in which the feedback mechanisms of one cluster are similar to each other while being dissimilar to others (Kaufman and Rousseeuw, 1990). We used the K-mode clustering method, since it is particularly suitable for non-numeric, categorial data (Chaturvedi et al., 2001). In order to get a statistical indication for a reasonable K (number of clusters), we calculated the the silhouette scores, a measure of cohesion and separation, for 2 to 10 clusters (Rousseeuw, 1987). Given the results, we inspected the four-cluster and five-cluster solutions that showed the highest silhouette scores. Due to practical and statistical judgement we chose the four-cluster solution with a score of 0.37, showing the most interesting and plausible findings and correlations (Balijepally et al., 2011). As **Table 2** of the following chapter depicts given the distribution of our samples, a number of relationships become apparent with four chatbot feedback clusters. First, the different characteristics under the feedback access dimension can be clearly classified as to whether they occur along active or passive

initiation. Additionally, the clusters are clearly split by feedback initiation and feedback access dimensions, as two clusters contain actively, the other two containing passively initiated feedback mechanisms.

4.4 Cluster Interpretation

Table 2. Frequency Distribution of Feedback Dimensions and Characteristics in 4 Clusters

		Amount per cluster			
	Amount	Response-Prompted Feedback	Conversation-Prompted Feedback	Implicit Optional Feedback	Explicit Optional Feedback
Characteristics	**82**	**29**	**19**	**9**	**25**
Feedback initiation					
Active	44	25	19	0	0
Passive	38	4	0	9	25
Feedback access					
End of conversation	23	4	19	0	0
End of dialog	21	21	0	0	0
Intent	10	1	0	9	0
Top-level item	15	0	0	0	15
In-chat button after response	10	3	0	0	7
Menu item	3	0	0	0	3
Feedback presentation					
In-chat buttons	26	22	1	2	1
Chat widget	8	3	3	0	2
Dialog	6	3	2	1	0
Pop-Up Window	28	0	11	0	17
External Link	14	1	2	6	5
Feedback structure					
Scale	73	29	19	1	24
Options	8	2	3	1	2
Free text	47	5	17	1	24
Unstructured	8	0	0	8	0
Pre-labeling of feedback					
Pre-labeling	61	19	11	8	23
No pre-labeling	21	10	8	1	2
Feedback specificity					
Message content	10	3	0	0	7
Technology	1	0	0	0	1
Service delivery	56	26	18	1	11
Company	9	0	0	8	1
Unspecified	6	0	1	0	5
Feedback scope					
Single response	10	3	0	0	7
Sub-dialog	20	20	0	0	0
Conversation	44	6	19	1	18
General	8	0	0	8	0

The four clusters gained trough our taxonomy are shown in **Table 2** in form of a sample distribution and are additionaly visualized as chatbot interface examples in Figure 3.

(1) The first cluster, the “Response-Prompted Feedback” archetype represents feedback mechanisms that are triggered when a dialog in the chatbot ends. This might be a response to a single question or the end of a multi-step dialog, but not the end of the entire conversation. The chatbot actively initiates the feedback process, e.g., by asking the user whether their goal was accomplished in a satisfactory manner. The user is provided with inputs in the form of buttons or scales that appear in the chat window among the other messages bubbles. The feedback is typically collected just in the chat interface. Due to the timing and the the questions, the feedback is supposed to be provided with regards to the most recent dialog. The feedback questions of most samples in this cluster also focussed on the users’ perceptions of the chatbot’s performance.

(2) The second cluster, the "Conversation-Prompted Feedback" archetype is initiated by the chatbot as well. However, these feedback mechanisms are triggered once the entire conversation and the service delivery is finished, e.g. when the customer closes the chat or via a timeout. Instead of presenting the feedback between the chat messages, the feedback mechanisms in this cluster usually opt for a presentation in a new window that is laid over the chat window, having the user focus solely on the feedback form. These forms usually have both a rating scale and one or two questions that users have to answer by writing texts. The feedback is framed to address the service delivery of the chatbot. Instead of focusing on a specific dialog, this mechanisms prompts the user to think about the overall chat and service experience and the entire conversation.

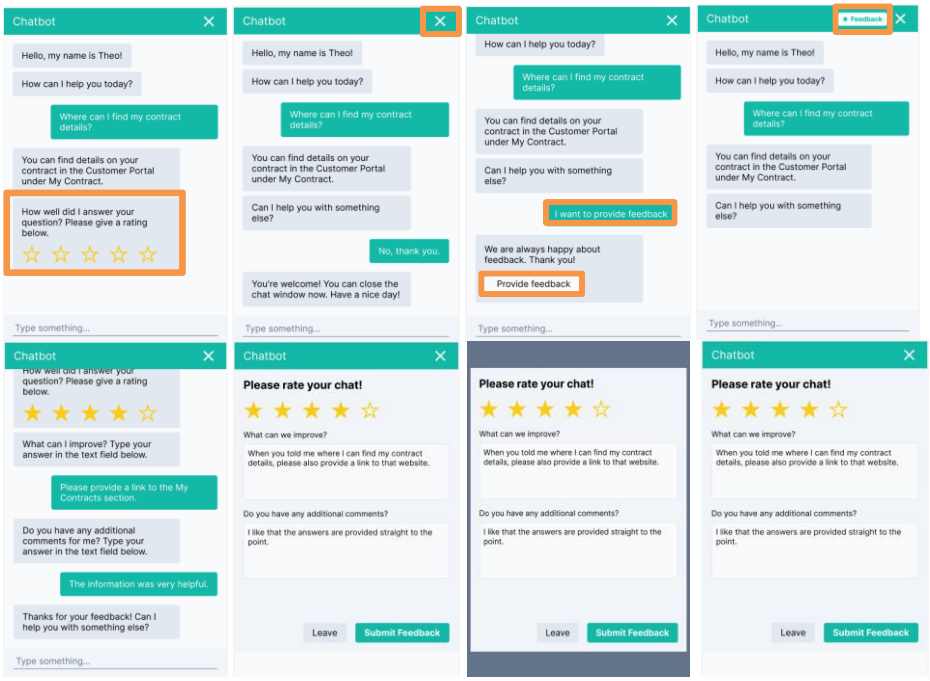

Figure 3. Exemplary Illustrations of the 4 Feedback Archetypes

(3) Unlike the previous two clusters, the feedback in the third cluster, the "Implicit Optional Feedback" is initiated by the user and not the chatbot (Sampson, 1998). These feedback designs are intent-based, meaning the chatbot uses NLU to detect intents or keywords like "feedback" or "complaint". Usually, when such an intent is detected, users receive a link to or can send an external contact form, e.g., the service company's general contact form. These forms often contain just a text input field for users to provide any feedback without being prompted to focus on the chatbot interaction. Accordingly, the feedback has a wider scope and is more related to the company.

(4) Lastly, the feedback mechanism of the fourth archetype, "Explicit Optional Feedback", is passive as well, meaning the user has to initiate the feedback process. This is done using a UI element, mostly located in the chatbot header, but also menu items or

buttons belong to this cluster. While they often trigger the same feedback presentation, a feedback form overlaying the chat window or an external link, they are less visually prominent while the top-level item encourages or at least reminds the user to provide optional feedback. The feedback structure usually contains scales and free text fields, asking users to rate their experience and provide additional comments. The users therefore typically refer more to the chat experience then to specific events or the company.

5 Discussion

5.1 Guidelines for Chatbot Feedback Mechanisms

In this chapter, we discuss the special features and possible applications of chatbot feedback mechanisms, both in general and with reference to the results of our taxonomy and clustering. We then summarize our contributions and identify limitations and further research opportunities in the area of chatbot feedback.

As literature has shown, providing feedback is valuable for chatbot improvement, but often very scarce in the case of chatbots (Akhtar et al., 2019). There is a general tendency for dissatisfied IS users to rather use (passive) feedback mechanisms to willingly provide feedback without being prompted (Baumeister et al., 2001; Nasr et al., 2014; Akhtar et al., 2019). This negative feedback is not reputationally damaging for service providers, as chatbot feedback is not publicly visible, but a bad chatbot is. On the other hand, non-public feedback means that the intrinsic motivation of feedback providers to help others is reduced (Hennig-Thurau et al., 2004). However, the provision of feedback can be increased by actively prompting for feedback like more than 50% of our inspected chatbots did via Response- or Conversation-Prompted Feedback. Although this can annoy customers and impair the user experience, given that they only have very sporadic and short interactions with the chatbot anyway (3-4 turns), this does not weigh as heavily as being annoyed when using software on a daily basis (Schuetzler et al., 2019). The deterrence of feedback prompts in chatbots can also be reduced, as recommended by research, by introducing the feedback gradually and partially optionally, for example in a small click dialog without blocking the input field, instead of overwhelming the user immediately (Han et al., 2021; Te Pas et al., 2020; Zierau et al., 2020). Nevertheless, when using active feedback mechanisms, chatbot managers should think about where feedback should be initiated. For example, the successful completion or the termination of pre-defined processes or dialogs is a good way to ask for feedback (Zierau et al., 2020). The end of a conversation, on the other hand, is always undefined within chatbots since users, aware of interacting with a machine, might simply close the browser tab instead of selecting a dedicated exit, as needed for the feedback mechanisms of cluster 2. Additionally, if, as 10 of the chatbots of our sample did, the feedback request is used undifferentiated, such as after every simple response (“Hello” – “Hi”), customers could be unnecessarily annoyed or the feedback could simply be completely unrelated to a real service task (Wang and Strong, 1996). However, this is where the potential of chatbot feedback can be realized. With in-app feedback mechanisms, a connection to an IS, i.e. a chabot response, a conversation or a service process can be

established, which is not possible with undifferentiated and general feedback forms (Van Oordt and Guzman, 2021). Therefore, chatbot managers should be clear about the goal they are pursuing with the feedback: Do they want to identify dissatisfied customers? Do they want an evaluation of their services or company? Or do they explicitly want to improve the chatbot? The more the latter applies, the more the feedback initiation and structure must be visually and temporally linked to the actions of the chatbot. A typical example of this are the thumbs up/down buttons, which in some cases were completely undifferentiated clickable and did not trigger any further actions. For deploying NLG, like it can be seen in ChatGPT, however, the evaluation of AI generated responses can be a valuable feedback (in the future) to evaluate (or compare) the quality of responses with real users (OpenAI, 2023; Ricciardelli and Biswas, 2019). However, such field test sets are more suited for chatbot developers than chatbot managers. If the latter are more interested in general feedback on the service and company, they can use the recognition of feedback intents (cluster 3) and the prominent placement of the passive feedback option (cluster 4). The specificity of the feedback received then depends above all on the feedback structure; scales and questions may narrow and guide the feedback, but also restrict the customer's expression (Haug et al., 2023; Van Oordt and Guzman, 2021). It is therefore advisable to combine easily analysable and relatable scales with free text. In general, we also found that some feedback mechanisms were mapped too little to the customer and too much to the chatbot developers, for example with regard to pre-labeling. Some chatbots suggested feedbacks such as "I was not understood" or "The answer did not help me", which are aimed at differentiating between NLU and knowledge base error, which the customer cannot necessarily follow.

5.2 Limitations and Future Research

In this paper we presented a taxonomy and a clustering that categorizes and explains feedback mechanisms of customer service chatbots based on literature and a real-world sample. We provide a theoretical structure for the knowledge on chatbot feedback and offer practical guidelines and orientation for feedback design, so chatbot managers can make informed decisions. Moreover, we equip researchers with a suitable framework for discussion and encourage for further research on chatbot feedback. Future research could broaden our context of task-oriented chatbots for customer service. Furthermore, we have focused on the specific channel of web- and text-based chatbots, while some of the findings of conversation-based feedback design apply to all media (Xiao et al., 2021). Finally, we have discussed the different feedback mechanisms only theoretically. Explicitly testing and comparing these feedback mechanisms in the field remains a a precious and exciting possibility for future research. On top of that, our taxonomy is an intentional simplification that does not capture all real world manifestations. For example, even the choice of texts or emojis in the feedback initiation or survey structure alone could create differences in perceptions and feedback outcomes (Alismail and Zhang, 2018). These or related aspects can be the subject of future research sheding light on the specifics of a chatbot feedback cluster or across different of our clusters.

References

Adamopoulou, E. and Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications* 2, p. 100006.

Akhtar, M., Neidhardt, J. and Werthner, H. (2019). The Potential of Chatbots: Analysis of Chatbot Conversations. In: *2019 IEEE 21st Conference on Business Informatics (CBI)*. Moscow, Russia: IEEE, pp. 397–404.

Alismail, S., Zhang, H. (2018). The Use of Emoji in Electronic User Experience Questionnaire: An Exploratory Case Study. In: *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Ashfaq, M., Yun, J., Yu, S. and Loureiro, S.M.C. (2020). I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telematics and Informatics* 54, p. 101473.

Balijepally, V., Mangalaraj, G., Iyengar, K. and Prairie, V. (2011). Are We Wielding this Hammer Correctly? A Reflective Review of the Application of Cluster Analysis in Information Systems Research. *Journal of the Association for Information Systems* 12(5), pp. 375–413.

Barki, H. and Hartwick, J. (1991). User participation and user involvement in information system development. In: *Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences*. Kauai, HI, USA: IEEE Comput. Soc. Press, pp. 487–492.

Baumeister, R., Bratslavsky, E., Finkenauer, C., De Vohs, K. (2001). Bad is Stronger than Good. Review of General Psychology 5(4), pp. 323-370.

Beaver, I. and Mueen, A. (2020). Automated Conversation Review to Surface Virtual Assistant Misunderstandings: Reducing Cost and Increasing Privacy. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(8), pp. 13140–13147.

Brandtzaeg, P.B. and Følstad, A. (2017). Why People Use Chatbots. In: Kompatsiaris, I. et al. eds. *Internet Science*. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 377–392.

Chaturvedi, A., Green, P.E. and Caroll, J.D. (2001). K-modes Clustering. *Journal of Classification* 18(1), pp. 35–55.

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering* 22(5), pp. 811–817.

Damodaran, L. (1996. User involvement in the systems design process-a practical guide for users. *Behaviour & Information Technology* 15(6), pp. 363–377.

De Keyser, A., Köcher, S., Alkire (née Nasr), L., Verbeeck, C. and Kandampully, J. (2019). Frontline Service Technology infusion: conceptual archetypes and future research directions. *Journal of Service Management* 30(1), pp. 156–183.

Fitzsimmons, J.A. and Fitzsimmons, M.J. (2011). *Service management: operations, strategy, information technology*. 7th ed. New York: McGraw-Hill.

Følstad, A. and Taylor, C. (2021). Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues. *Quality and User Experience* 6(1), p. 6.

Fricker, R.D. and Schonlau, M. (2002). Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature. *Field Methods* 14(4), pp. 347–367.

Gao, M., Liu, X., Xu, A. and Akkiraju, R. (2021). Chatbot or Chat-Blocker: Predicting Chatbot Popularity before Deployment. In: *Designing Interactive Systems Conference 2021*. Virtual Event USA: ACM, pp. 1458–1469.

Glaser, B.G. and Strauss, A.L. (1999). *The discovery of grounded theory: strategies for qualitative research*. London New York: Routledge.

Glass, R.L. and Vessey, I. (1995). Contemporary application-domain taxonomies. *IEEE Software* 12(4), pp. 63–76.

Gopinath, K. and Kasilingam, D. (2023). Antecedents of intention to use chatbots in service encounters: A meta-analytic review. *International Journal of Consumer Studies*.

Gould, J.D. and Lewis, C. (1985). Designing for usability: key principles and what designers think. *Communications of the ACM* 28(3).

Grudin, J. and Jacques, R. (2019). Chatbots, Humbots, and the Quest for Artificial General Intelligence. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow Scotland Uk: ACM, pp. 1–11.

Han, X., Zhou, M., Turner, M. and Yeh, T. (2021). Designing Effective Interview Chatbots: Automatic Chatbot Profiling and Design Suggestion Generation for Chatbot Debugging. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. pp. 1–15.

Harris, M.A. and Weistroffer, H.R. (2009). A New Look at the Relationship between User Involvement in Systems Development and System Success. *Communications of the Association for Information Systems* 24.

Haug, S., Benke, I. and Maedche, A. (2023). Aligning Crowdworker Perspectives and Feedback Outcomes in Crowd-Feedback System Design. *Proceedings of the ACM on Human-Computer Interaction* 7(CSCW1), pp. 1–28.

Hennig-Thurau, T., Gwinner, K.P., Walsh, G. and Gremler, D.D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing* 18(1), pp. 38–52.

Janssen, A.H.A., Grützner, L. and Breitner, M.H. (2021). Why do Chatbots fail? A Critical Success Factors Analysis, p. 18.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. 1st ed. Wiley.

Kujala, S. (2003). User involvement: A review of the benefits and challenges. *Behaviour & Information Technology* 22(1), pp. 1–16.

Kvale, K., Freddi, E., Hodnebrog, S., Sell, O.A. and Følstad, A. (2021). Understanding the User Experience of Customer Service Chatbots: What Can We Learn from Customer Satisfaction Surveys? In: Følstad, A., Araujo, T., Papadopoulos, S., Law, E. L.-C., Luger, E., Goodwin, M., and Brandtzaeg, P. B. eds. *Chatbot Research and Design*. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 205–218.

Lewandowski, T., Heuer, M., Vogel, P., and Böhmann, T. (2022). Design knowledge for the lifecycle management of conversational agents, 17. Internationale Tagung Wirtschaftsinformatiok, Nürnberg. Proceedings.

Li, T.J.-J., Chen, J., Xia, H., Mitchell, T.M. and Myers, B.A. (2020). Multi-Modal Repairs of Conversational Breakdowns in Task-Oriented Dialogs. In: *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. Virtual Event USA: ACM, pp. 1094–1107.

Lin, W.-C., Tsai, C.-F., Hu, Y.-H. and Jhang, J.-S. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences* 409–410, pp. 17–26.

Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P. and Theeramunkong, T. (2019). A Survey on Evaluation Methods for Chatbots. In: *Proceedings of the 2019 7th International Conference on Information and Education Technology*. Aizu-Wakamatsu Japan: ACM, pp. 111–119.

McKeen, J.D. and Guimaraes, T. (1997). Successful Strategies for User Participation in Systems Development. *Journal of Management Information Systems* 14(2), pp. 133–150.

Meyer-Waarden, L., Pavone, G., Poocharoentou, T., Prayatsup, P., Ratinaud, M., Tison, A. and Torné, S. (2020). How Service Quality Influences Customer Acceptance and Usage of Chatbots? *Journal of Service Management Research* 4(1), pp. 35–51.

Nasr, L., Burton, J., Gruber, T. and Kitshoff, J. (2014). Exploring the impact of customer feedback on the well-being of service entities: A TSR perspective. Bo Edvardsson And Professor Anders Gustafsson, P. ed. *Journal of Service Management* 25(4), pp. 531–555.

Nickerson, R.C., Varshney, U. and Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems* 22(3), pp. 336–359.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In: *CHI '94: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Ordenes, F.V., Theodoulidis, B., Burton, J., Gruber, T. and Zaki, M. (2014). Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach. *Journal of Service Research* 17(3), pp. 278–295.

OpenAI. (2023). https://chat.openai.com/chat (visited on November 10, 2023).

Pagano, D. and Bruegge, B. (2013). User involvement in software evolution practice: A case study. In: *2013 35th International Conference on Software Engineering (ICSE)*. San Francisco, CA, USA: IEEE, pp. 953–962.

Pagano, D. and Maalej, W. (2013). User feedback in the appstore: An empirical study. In: *2013 21st IEEE International Requirements Engineering Conference (RE)*. Rio de Janeiro-RJ, Brazil: IEEE, pp. 125–134.

Poblete, B. and Baeza-Yates, R. (2008). Query-sets: using implicit feedback and query patterns to organize web documents. In: *Proceedings of the 17th international conference on World Wide Web*. Beijing China: ACM, pp. 41–50.

Ricciardelli, E. and Biswas, D. (2019). Self-improving Chatbots based on Reinforcement Learning. Montreal, Canada.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, pp. 53–65.

Sampson, S.E. (1998). Gathering customer feedback via the Internet: instruments and prospects. *Industrial Management & Data Systems* 98(2), pp. 71–82.

Schuetzler, R.M., Grimes, G.M., Giboney, J.S. and Rosser, H.K. (2021). Deciding Whether and How to Deploy Chatbots. *MIS Quarterly Executive*, pp. 1–15.

Sommerville, I. (2011). *Software engineering*. 9th ed. Boston: Pearson.

Strauss, A.L. and Corbin, J.M. (1998). *Basics of qualitative research: techniques and procedures for developing grounded theory*. 2nd ed. Thousand Oaks: Sage Publications.

Strohmann, T., Khosrawi-Rad, B., Schmidt, L. and Hiske, P. (2023). AI-based Technologies for Conversational Agent Design– Development Tools and Architectures for Intelligent Interactions. Panama City, Panama.

Te Pas, M.E., Rutten, W.G.M.M., Bouwman, R.A. and Buise, M.P. (2020). User Experience of a Chatbot Questionnaire Versus a Regular Computer Questionnaire: Prospective Comparative Study. *JMIR Medical Informatics* 8(12), p. e21982.

Tizard, J., Rietz, T. and Blincoe, K. (2020). Voice of the Users: A Demographic Study of Software Feedback Behaviour. In: *2020 IEEE 28th International Requirements Engineering Conference (RE)*. Zurich, Switzerland: IEEE, pp. 55–65.

Van Doorn, J., Lemon, K.N., Mittal, V., Nass, S., Pick, D., Pirner, P. and Verhoef, P.C. (2010). Customer Engagement Behavior: Theoretical Foundations and Research Directions. *Journal of Service Research* 13(3), pp. 253–266.

Van Oordt, S. and Guzman, E. (2021). On the Role of User Feedback in Software Evolution: a Practitioners' Perspective. In: *2021 IEEE 29th International Requirements Engineering Conference (RE)*. Notre Dame, IN, USA: IEEE, pp. 221–232.

Wang, R.Y. and Strong, D.M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems* 12(4), pp. 5–33.

Wirtz, J. and Tomlin, M. (2000. Institutionalising customer-driven learning through fully integrated customer feedback systems. *Managing Service Quality: An International Journal* 10(4), pp. 205–215.

Witell, L., Kristensson, P., Gustafsson, A. and Löfgren, M. (2011). Idea generation: customer co-creation versus traditional market research techniques. *Journal of Service Management* 22(2), pp. 140–159.

Xiao, Z., Mennicken, S., Huber, B., Shonkoff, A. and Thom, J. (2021). Let Me Ask You This: How Can a Voice Assistant Elicit Explicit User Feedback? *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2), pp. 1–24.

Zierau, N., Hausch, M., Bruhin, O. and Söllner, M. (2020). Towards Developing Trust-Supporting Design Features for AI-Based Chatbots in Customer Service. *ICIS Proceedings*, pp. 1-9.